ChinaXiv合作期刊

基于维基百科的中文文本层次路径生成研究*

夏　天
(中国人民大学数据工程与知识工程教育部重点实验室　北京　100872)
(中国人民大学信息资源管理学院　北京　100872)

摘要: 【**目的**】利用维基百科知识库生成自由文本的层次语义路径。【**方法**】针对维基百科的中文导出数据，构建层次结构的树状图；进而通过显性语义分析将自由文本表示为文章概念向量，通过文章–类别关联关系将文本映射到树状图中构成种子类别节点，再通过种子节点开始的信息扩散和自顶向下的路径选择与优化，生成层次路径。【**结果**】首条层次路径的平均相关度在测试集上达到 54.10%，前 20 条路径整体上按相关度降序排序。【**局限**】未分析显性概念向量在保留不同概念数量时对生成路径质量的影响。【**结论**】基于维基百科知识库所生成的层次路径结果能够反映文本的主要语义信息。

关键词: 语义路径　显性语义分析　层次分类　维基百科
分类号: G353

chinaXiv:201711.01237v1

1　引　言

文本的语义描述是文本分析的常见任务，根据描述粒度的不同可以分为三个层次：以词袋法为主的细粒度表示，将文本看作是由相互独立且具有不同权重的词语构成的集合，权重计算有布尔逻辑、TF-IDF 等方法；以分类为代表的粗粒度表示，通过构建朴素贝叶斯、SVM、决策树等分类模型，自动从预定义的类别集合中选择最相关的分类；介于前二者之间的描述方式，以图结构和主题模型最为常见，前者把文本表示为由概念节点及关联边构成的语义图[1]，后者以 LDA 为典型代表[2]，把文本看作是由若干个主题按照某种分布生成的结果，主题本身又是由词语根据特定分布生成。

在三种不同粒度的处理方式中，分类对于文本的语义描述最为概括，人工可读性最强。然而，传统分类技术所处理的类别集合数量固定，各分类之间在语义上处于相同等级，不存在上下位层次关系，无法深度刻画文本的语义信息，如能引入多级分类，通过带层次结构的语义路径对文本进行描述，将有利于更好地快速获取文本的主要语义。

因此，本文围绕如何识别自由文本的层次语义路径进行研究，基于维基百科中文导出数据，构建了带有大规模层级结构的树状图，借助显性语义分析将任意文本的语义信息映射到树状图中，进而通过节点信息扩散和路径求解与优化，生成文本对应的层次分类路径。

2　相关工作

文本的层次语义描述可以借助层次分类实现，即按照一个规模巨大的类别层次，指定未知对象在层次中所隶属的类别[3]。层次分类需要良好的层次结构和

通讯作者：夏天，ORCID：0000-0001-7564-7368，E-mail：xiat@ruc.edu.cn。
*本文系北京高等学校青年英才计划项目“基于链接和主题分析的微博社区挖掘研究”(项目编号:YETP0215)和国家社会科学基金重大项目“国家数字档案资源整合与服务机制研究”(项目编号:13&ZD184)的研究成果之一。

chinaXiv:201711.01237v1

一定规模的训练数据，通过将层次分类问题转换为传统分类，再利用常规分类算法实现[4-5]。然而，人工维护一棵组织严谨的大规模层次树难度较大，分类节点数量众多使得分类算法的效率较低，从而限制了层次分类的应用范围。

相比层次分类而言，利用维基百科现有的文章和分类网络识别文本的层次语义更具优势：一方面，维基百科已经形成了开放的、动态增长的分类体系；另一方面，维基百科的分类与文章之间的链接引用关系提供了更多的显性语义信息，在此之上已有部分较为有效的文本语义分析技术[6]。

其中，Muchnik 等[7]利用维基百科的文章链接网络，自动构建术语在网络中的潜在层次结构，但该研究未使用维基百科的分类信息。Gabrilovich 等[8]提出的显性语义分析(Explicit Semantic Analysis，ESA)是基于维基百科的文本语义表示的经典方法，该方法使用维基百科的文章及其之间的链接信息，把文本表示为由概念(文章标题)构成的向量，在词语相关度计算[9]、查询扩展[10]、文本分类[8]等应用中得到了广泛应用。ESA 表达的是文本与维基概念之间在统计意义上的相关性，概念向量中的各元素之间与词袋法一样维持了独立性假设，因此，ESA 对文本实际语义的直观解释能力依然较弱，以本文所用数据集构建的 ESA 模型和待分析文本 “新浪微博” 为例，ESA 输出的前 5 篇最相关文章分别为 “腾讯微博”、“长微博”、“微博 AIR”、“自由微博” 和 “对新浪微博的争议”，而通过本文的路径识别技术所输出的前两个层次路径为 “社会/大众媒体/全球资讯网/Web2.0” 和 “社会/文化/网络文化/虚拟社群”，显然，层次路径更能准确描述文本的语义信息。

总体而言，借助于词条概念描述文本的语义已有较好的研究进展，但如何生成任意文本的层次语义路径尚无公开有效的方法。

本文直接面向开放的维基百科分类体系，在由文章与分类、分类之间共同构成的巨大网络中，抽取可表达文本主要语义的层次路径。与层次分类法不同，本文方法所处理的分类数量巨大，分类网络复杂；路径识别不需要构建复杂的分类器，而是在简化网络中借助于信息扩散动态完成，最终生成可读性强的层次语义路径。

3 基于维基百科的层次分类树状图构建

维基百科提供了开放的层次分类体系，用于对以文章为单位概念的世界知识进行多维度标注，中文分类形式上以杜威十进位图书分类法为主，同时参考《中国图书馆分类法》以及赖永祥《中国图书分类法》。维基百科数据由人工编纂而成，凝聚了群体智慧，内容相对丰富完整，并且可自由获取使用，因此，笔者采用维基百科数据构建大规模层次分类体系。

维基百科分类体系可以表示为由节点集 $V=\{v_1,v_2,\cdots,v_n\}$ 和弧集 $E=\{e_1,e_2,\cdots,e_m\}$ 组成的有向图 $G_W=<V,E>$，其中，节点 $v_i(1\leqslant i\leqslant n)$ 表示类别；弧 $e_i=<v_j,v_k>(1\leqslant i\leqslant m,1\leqslant j,k\leqslant n)$表示类别节点 v_j 是 v_k 的父类别，维基百科分类图 G_W 的一个子图如图 1 所示：

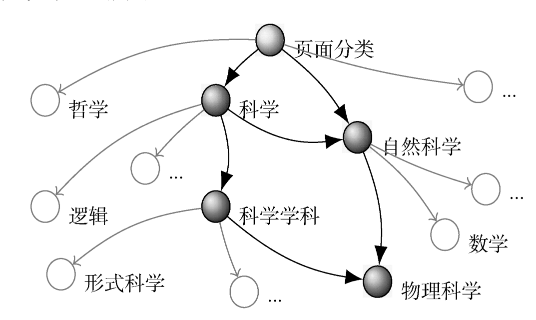


图 1　维基百科分类图子图

维基百科的中文分类体系以 “页面分类” 为总入口，该分类下拥有 22 个直接子分类，也是维基百科有实际意义的第一级分类，如表 1 所示：

表 1　维基百科的第一级分类列表

序号	名称	序号	名称
1	哲学	12	心理学
2	人物	13	科技
3	历史	14	资讯
4	宗教	15	跨学科领域
5	文学	16	休闲
6	艺术	17	人文学科
7	自然科学	18	应用科学
8	社会科学	19	社会
9	地理	20	技术
10	科学	21	总类
11	语言	22	词汇列表

ChinaXiv合作期刊

chinaXiv:201711.01237v1

为便于描述，对 G_W 做如下设定:

(1) 令"页面分类"为 G_W 的根节点，记为 $root(G_W)$。

(2) 对于一条弧 $e = <v_i, v_j>$，称 v_i 是 v_j 的父节点，v_j 是 v_i 的子节点，令 parents(v) 表示 v 的所有父节点集合，children(v) 表示 v 的所有子节点集合。例如，图 1 中有:

children("自然科学")={ "物理科学", "数学"…}
parents("物理科学")={ "科学学科", "自然科学"…}

(3) 对每个节点 v 赋予一个相对于根节点位置的深度属性 depth，当 v 是根节点时，深度为 0，其他情况递归定义如下:

$$depth(v) = \min_{v_i \in parents(v)}(depth(v_i) + 1) \qquad (1)$$

例如，在图 1 中，有:

depth("科学")=depth("自然科学")=1
depth("科学学科")=depth("物理科学")=2

(4) 将从第一级节点开始到类别节点 v 为止的任一条简单路径称为 v 的一条层次分类路径，简称路径，记为 p_v，并令 $|p_v|$ 表示路径的长度，即 p_v 所包含的分类节点数量，如图 1 中，"自然科学→物理科学"是节点"物理科学"的一条分类路径，其长度为 2。

完整的维基百科分类图 G_W 存在许多不利于算法自动分析的方面：首先，维基百科拥有大量以不同侧面对概念进行分类的情况，如"总类→分类→直接命名的分类→以人物命名的分类→以各职业人物命名的分类→以商人命名的分类→比尔·盖茨"，去除这些以导航为主要目的的分类，可以提高路径自动识别的效果。其次，经拓扑排序发现图 G_W 中存在大量环路，不利于分类体系的递归处理，如"社会科学→刑事学→罪案→侵犯人权→宗教迫害→宗教多元主义→宗教迫害"。再次，部分节点存在路径包含现象，该现象是指节点有两条以上长度不等的路径，并且长路径包含了短路径的所有类别，如图 1 中，路径"科学→自然科学→物理科学"和"自然科学→物理科学"均为"物理科学"的路径，通常情况下，仅保留短路径不破坏分类体系的主要语义信息，并能简化图的复杂程度。另外，G_W 中存在分类引用缺失、无有效路径以及类别重复等少量异常现象。例如，实验数据中的分类"佛教法器"，其父类指向了并不存在的"法器"分类；类别"各类型智慧"没有父分类，即不存在有效路径；"俄罗斯探险家"和"俄羅斯探險家"表达了相同的意义，但却对应两个完全独立的分类页面。

为解决以上问题，笔者提出了层次分类图构建算法，通过对 G_W 进行剪枝，移除部分节点和边，消除环路和路径包含现象，得到简化后的层次分类图 G_H，算法如下所示:

```
输入: 原始维基百科分类关系图 G_W
输出: 用于层次路径识别的树状图 G_H
1: R = root(G_W)
2: V_H = {R}, E_H = ø;
3: Init queue Q and Enqueue R into Q;
4: while Q not empty do
5:     v = dequeue from Q;
6:     if  v∈{"跨领域学科", "总类", "词汇列表"} then continue;
7:     for each child in children(v) do
8:         if depth(child) = depth(v) + 1 then
9:             V_H = V_H ∪ child;
10:            E_H = E_H ∪ edge(v → child);
11:            Enqueue child into Q;
12:        end if
13:    end for
14: end while
15: return G_H =< V_H, E_H >
```

算法借助于队列结构自根节点对 G_W 进行广度优先遍历，对于当前被访问的节点 v(行 5)，通过忽略"跨学科领域"、"总类"和"词汇列表"三个一级类别节点以消除侧面分类(行 6)；然后处理 v 的每一个子节点 child，当其深度为 v 的深度值加 1 时，把子节点 child 和边 v→child 分别加到节点集 V_H 和边集 E_H 中 (行 9，行 10)，否则，说明 G_W 中有除 v 之外的节点指向 child，且距离根节点更近，此时忽略边 v→child。最终，V_H 和 E_H 分别保存了精简后的层次分类图的节点集和边集，共同构成了的树状图 G_H。

算法保证了 G_H 拥有一个无入边的根节点，且图中每一个节点 v 的入边只来自上层节点，它们的深度为 v 的深度减 1，出边只指向下层节点，深度为 v 的深度加 1。G_H 具有树结构的多数特性：拥有根节点、子节点、叶子节点和分层结构，但 G_H 中节点的父节点不唯一，所以称之为树状图(Tree like Graph)。

4　层次分类路径识别方法

基于维基百科的语义层次分类路径识别分为三个部分：将自由文本表示为由维基文章构成的显性概念；

ChinaXiv合作期刊

chinaXiv:201711.01237v1

将显性概念映射到树状图并求解层次分类路径集合；综合考虑相关度和新颖度，优化层次分类路径的选择。

4.1 文本的显性概念表示

ESA 借助通用知识库，将自由文本表示为一组由概念构成的向量，通常采用维基百科训练得到。给定一组概念（对应于维基百科的文章标题）集合 $\{a_1,a_2,\cdots,a_n\}$ 和与之关联的文档(即维基百科文章的内容 $\{d_1,d_2,\cdots,d_n\}$，ESA 模型构造一个稀疏矩阵 T，其中每一列表示一个概念，每一行对应于一个出现在 $\bigcup_{i=1\cdots n} d_i$ 中的词语，T 中的每个元素 T[i,j]对应于出现在文档 d_j 中的词项 t_i 的 TF-IDF 值[8]：

$$T[i,j]=tf(t_i,d_j)\cdot\log\frac{n}{df_i} \tag{2}$$

如文献[8]所述，并非所有的文档对于ESA都有相同的效果，笔者从内容和链接关系两个方面对维基百科的原始文章进行过滤。在内容方面，如文章 a 是跳转页面、消歧页面、列表页面，或者文章 a 所包含的词语数量少于 200，则作为非重要文章予以过滤；在链接关系方面，如果文章 a 的出入链之和小于 20，则予以过滤。

为建立 ESA 模型，笔者对过滤后的维基百科数据进行扫描，计算每对“词语→文章”的 TF-IDF 值，形成最终的 ESA 矩阵 T，并进一步维护了文章到类别的隶属关系，用于后续的种子类别选取，从而构成了自由文本到层次路径之间的桥梁关系。

在构建矩阵 T 之后，给定文本 t，其显性语义概念向量可由以下公式计算得到：

$$\overline{V}_t=\sum_{w_i\in terms(t)} tf(w_i,t)\cdot idf(w_i)\cdot T[w_i] \tag{3}$$

其中，$tf(w_i,t)$ 表示词语 w_i 在文本 t 中的词频，$idf(w_i)$ 表示 w_i 在所有维基百科数据集上的倒排文档频率，$T[w_i]$ 表示矩阵 T 中 w_i 所对应的行向量，即其显性语义向量。

为获取文本的主要语义概念，笔者对向量 $\overline{V}_t$ 按照其元素得分进行降序排序，并挑选前 n 个元素作为文本最终的显性语义分析结果，形式化表示为 $ESA_t=\{p(a_1|t),p(a_2|t),\cdots,p(a_n|t)\}$，$p(a_i|t)$ 表示文章 a_i 与文本 t 的语义相关程度。

4.2 分类节点的语义关联与扩散及分类路径求解

令 $CS(a_i)$ 表示文章 a_i 所隶属的关联分类，对于文本 t 和 G_H 中的分类 c，如存在 $a_j\in ESA_t$，使得 $c\in CS(a_j)$，则称分类 c 为文本 t 在树状图 G_H 上的初始种子类别节点，简称种子节点。令 $w(c_j|t)$ 表示种子节点 c_j 与文本 t 相关程度的权重大小，计算公式如下：

$$w(c_j|t)=\sum_{a_i\in c_j}\frac{p(a_i|t)}{|CS(a_i)|} \tag{4}$$

其中，$a_i\in c_j$ 表示文章 a_i 隶属于类别 c_j；$|CS(a_i)|$ 表示 a_i 所关联的分类集合的大小。公式(4)表明，种子节点的分类权重由所关联文章的 ESA 分值按比例累加得到。进一步，为保证所有种子节点的权重之和为 1，笔者对种子节点的权重进行归一化处理，记为 $w'(c_i|t)$，公式如下：

$$w'(c_i|t)=\frac{w(c_i|t)}{\sum_{c_j\in\Delta} w(c_j|t)} \tag{5}$$

其中，Δ表示所有种子节点集合。为保持定义的完整性，如果 $c_i\notin\Delta$，令 $w'(c_i|t)=0$。

做如下假设：任一文本 t 均可以由维基百科的分类加以描述，描述时由根节点开始，经中间层级的分类自顶向下逐级细化描述，直至到达维基编纂人员认可的细粒度分类为止；令 $p(c_i|t)$ 表示分类 c_i 与文本 t 的语义相关程度，对每一个分类进行相关度赋值后，就可以按照一定的策略从图中挑选出自根节点达到终止节点的最相关路径，作为文本 t 的语义层次路径。

在表现形式上，人们仅观察到与文章直接关联的种子分类节点，而自根节点到达种子节点所经过的中间节点隐藏在层次树 G_H 中，为求解中间节点及其相关度值，笔者提出反向扩散方法，自种子节点开始，将每个节点的相关度值向父节点扩散，直至根节点为止，此时有 $p(root(G_h)|t)=1$，即所有种子节点的信息最终汇集到根节点，任一文本均隶属于维基百科的根分类。

令 $I(c_j\to c_i|t)$ 表示节点 c_j 扩散到节点 c_i 的信息量，定义如下：

$$I(c_j\to c_i|t)=\frac{p(c_j|t)\cdot count(c_i)}{\sum_{c_k\in parents(c_j)} count(c_k)} \tag{6}$$

其中，$count(c_i)$ 表示隶属于节点 c_i 或 c_i 子孙节点的文章数量。此时，$p(c_i|t)$ 求解如下：

$$p(c_i|t)=w'(c_i|t)+\sum_{c_j\in children(c_i)} I(c_j\to c_i|t) \tag{7}$$

ChinaXiv合作期刊

chinaXiv:201711.01237v1

即节点 c_i 与文本 t 的语义相关度由直接关联的文章所传递的 ESA 权重和所有子节点所传递的信息量共同决定。从种子节点开始，自底向上依次计算，即可求解所有中间节点及根节点与文本 t 的语义相关度值。然后，从根节点开始，通过所有相关度大于 0 的中间节点，到种子节点为止，即可获取到所有可能的分类路径。对于任一分类路径 $path_k =< c_1, c_2, \cdots, c_k >$，定义其与文本 t 的语义相关度 $PR(path_k)$ 如下:

$$PR(path_k \mid t) = \frac{\sum_{c_i \in path_k} p(c_i \mid t)}{\mid c_i \in path_k \mid} \tag{8}$$

根据公式(8)对每条层次分类路径按其关联度由高到低排序，并挑选得分最高的前 N 个作为候选结果，即可实现对文本 t 的语义路径识别。

4.3　层次分类路径的优化选择

为保证生成路径的新颖性与多样性，笔者参考文献[11]提出的方法对候选路径进行剪枝，移除高度相似的重复路径。首先，基于文本 t 的候选路径集，构建无向带权图 $G_I =< V_I, W_I, E_I >$，其中，G_I 的每个节点 $v_i \in V_I$ 对应于一条分类路径 $path_i$，其权重 $w_i \in W_I$ 为 $PR(path_i \mid t)$。对于任意两个节点 v_i、v_j 及其对应的层次分类路径 $path_i$、$path_j$，如 $path_i$ 与 $path_j$ 的相似度 $sim(path_i, path_j)$ 大于指定阈值，则图 G_I 存在无向边 $e = (v_i, v_j) \in E_I$。

根据如下贪心策略挑选独立路径:

(1) 从图 G_I 中选取权重最大的节点 v 作为有效路径并予以标记，删除与 v 相邻的节点及边，并把 v 添加到队列 Q 的尾部;

(2) 重复以上过程直至图 G_I 中的所有节点被挑选或删除完毕。

此时，队列 Q 中保存了所有互不依赖的层次路径，并按照语义相关度由高到低排列。

在上述步骤中，如何计算任意两路径之间的相似度至关重要，笔者采用如下方式:

$$sim_p(p_i, p_j) = \frac{\sum_{k=1}^{L} (L-k+1) \cdot sim_c(c(p_i,k), c(p_j,k))}{\sum_{k=1}^{L} k + (\| p_i \| - \| p_j \|)} \tag{9}$$

$L = \min(|p_i|, |p_j|)$，$c(p_i, k)$ 表示路径 p_i 中的第 k 个类别，$sim_c(c_1, c_2)$ 表示两个类别 c_1 和 c_2 的相似度，为简化复杂度，令 $c_1 = c_2$ 时，$sim_c(c_1, c_2) = 1$，否则为 0。

5　实　验

为验证本文方法的效果，笔者构建了维基百科训练数据集和测试数据集，以算法生成的层次分类路径有序列表作为测试对象，对比生成路径和文章自带的原始类别的相关度，以反映自动生成路径的实际效果。

5.1　实验数据

选取维基百科 2015 年 6 月发布的中文导出数据 “zhwiki-20150602-pagesarticles-multistream.xml.bz”①，该数据集共包含 2 648 029 个页面，其中，文章页面占 55.93%，分类页面占 7.47%，文档附件、图片等其他类型资源页面占 36.60%，具体组成如表 2 所示。通过数据清洗处理，最终保留了 184 968 个文章页面和 176 484 个分类页面，分别用于构建 ESA 模型和层次分类树状图。

表 2　维基百科实验数据集页面组成情况

页面类型	数量	百分比	子类型	数量	百分比
文章页面	1 480 963	55.93%	跳转文章数量	658 084	44.44%
			内容过滤数量	610 183	41.20%
			链接过滤数量	27 728	1.87%
			有效文章数量	184 968	12.49%
分类页面	197 872	7.47%	特殊分类数量	21 304	10.77%
			简繁同名异常数量	84	0.04%
			有效分类数量	176 484	89.19%
其他页面	969 194	36.60%	—	—	—

清洗后的维基百科分类图 G_W 共拥有 176 484 个节点和 335 329 条边，通过树状图构建算法进行过滤处理，去除环路和孤立点后，形成最终的层次分类图 G_H，包含 171 681 个节点和 220 861 条边，分别为 G_W 的 97.28%和 65.86%，即 G_H 基本保留了原图的分类名称，但去除了大量冗余路径。

为构建测试集，笔者从维基百科原始数据中去除 184 968 条训练数据，从剩余的 637 911 个非跳转文章

① http://dumps.wikimedia.org/zhwiki/20150602/.

页面中以 1.5%的概率随机抽样，去除字数少于 50 的文章，最终构成了包含 6 629 个文章的测试集[①]，测试集以 XML 格式保存，每个文章包含页面 Id、标题、去除标签后的文本和所隶属的分类。

5.2 实验数据

给定测试数据集中的一个文章 a_i，a_i 在维基百科原始数据中所隶属的分类集合为 $cs(a_i)$，通过本文方法计算得到的层次分类路径集合为 $PS(a_i)=\{path_1,path_2,\cdots,path_n\}$，定义 a_i 与任一条分类路径 $path_j$ 的相关度 R 如下：

$$R(a_i,path_j)=\max_{c\in CS(a_i)} rel(path_j,c) \tag{10}$$

其中，$rel(path_j,c)$ 表示类别 c 与给定路径 $path_j$ 的相关度，计算公式如下：

$$rel(path_j,c)=\frac{mnp(path_j,c)}{mnp(path_j,c)+dis(path_j,c)} \tag{11}$$

其中，$dis(path_j,c)$ 表示类别节点 c 在维基百科分类图中到达路径 $path_j$ 任一节点的最短距离，$mnp(path_j,c)$ 表示节点 c 与 $path_j$ 的距离取最小值时，在 $path_j$ 中相对应的匹配节点位置(Matched Node Position)。

进一步，令 $R(a_i,k)$ 表示文章 a_i 与计算得到的前 k 条层次分类路径的平均相关度，简记为 R@k，计算公式如下：

$$R(a_i,k)=\frac{1}{k}\cdot\sum_{j=1}^{k} rel(a_i,path_j) \tag{12}$$

5.3 实验结果与分析

取测试文章的标题和正文文本的前 300 个汉字作为自由文本，计算其显性概念向量，保留前 20 个主要概念用于生成种子分类节点，生成层次语义路径集合。为便于获得路径识别的感性认识，下面给出了测试集中的"中国古典典籍"和"邻接矩阵"两篇文章人工给出的分类信息和自动识别出的前 5 条路径，以及每条路径与文章的相关度 R 和 k-平均相关度(k∈[1,5])，如表 3 所示。

由表 3 可看出，本文所提方法能够从层次分类知识体系中对文本内容进行合适的语义定位，所输出的

表 3 层次路径识别结果示例

测试文章	原始类别	前 5 条识别路径	R	R@k
中国古典典籍	中国古典典籍	(1) 文学/文学体裁/语录	0.250	0.250
	中国思想	(2) 历史/历史学/文献学	0.400	0.325
	—	(3) 社会/教育/学术/反智主义	0.667	0.439
	—	(4) 历史/各种主题的历史/思想史/中国思想史	0.800	0.529
	—	(5) 社会/文化/各国文化/中国文化/经学	0.800	0.583
邻接矩阵	矩阵	(1) 自然科学/数学/离散数学/图论	1.000	1.000
	图论	(2) 应用科学/计算机科学/数据结构	1.000	1.000
	数据结构	(3) 应用科学/资讯科学/生物信息学	0.333	0.778
	—	(4) 资讯/信息论/编码理论	0.600	0.733
	—	(5) 应用科学/应用数学/数值分析/数值线性代数	0.333	0.653

层次路径能够从不同侧面反映文本的主要语义信息，与人工标注的细粒度分类具有较高的关联关系，能够为文章编纂人员对文本进行合理分类提供有效的参考借鉴。

为反映整体情况，根据公式(12)计算前 k 条路径与测试文章自带分类的k-平均相关度，k取不同数值时的实验结果如图 2 所示：

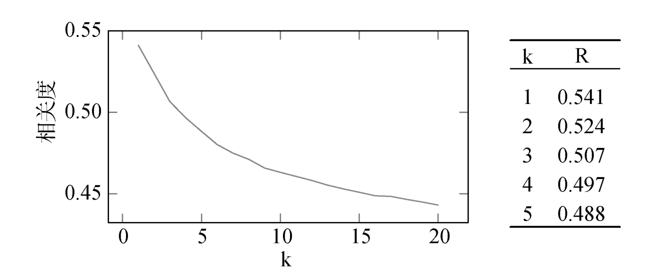


图 2 k-平均相关度实验结果 (k∈[1,20])

图 2 的右侧给出了 k 取值从 1 到 5 时，在整个测试数据集上的 R@k，左侧曲线则给出了 k 取值在 1 到 20 之间的整体变化情况。相关度均值随着 k 的增大而显著降低，说明识别结果整体上能够按照与原始文本

① https://github.com/iamxiatian/data/blob/master/zh.wiki6629.zip.

chinaXiv:201711.01237v1

ChinaXiv合作期刊

chinaXiv:201711.01237v1

的语义相关度由高到低排序；当k = 1时，平均相关度值达到0.541，则表明超过一半的情况下首条路径与人工标注的类别保持一致。部分测试文章的相关度较低的原因，一方面是由于方法本身和数据质量的局限，采用显性语义分析表示自由文本会引入噪声，另一方面则是人工标记的分类不够全面(见表3)，使得有较高语义相关度的路径在测试中的实际得分较低。

6　结　语

本文提出了一种基于维基百科的语义层次路径识别方法，该方法首先利用显性语义分析技术将自由文本表示为维基百科词条概念向量，进而通过词条与类别之间的隶属关系，将其关联到层次分类树状图之中，通过自种子分类节点向根节点的语义扩散和自顶向下的分类路径求解与优化，实现了对任意文本的语义层次路径标记。实验结果表明本方法自动生成的路径与人工标记的类别具有较高的关联度。

下一步研究包括:

(1) 探索新的分类节点在图中的信息扩散计算方式，进一步提高层次路径识别效果；

(2) 层次路径识别技术在相似度计算和分类等文本挖掘任务当中的应用。

参考文献:

[1] 吴江宁，刘巧凤. 基于图结构的中文文本表示方法研究[J]. 情报学报，2010，29(4): 618-624. (Wu Jiangning, Liu Qiaofeng. Research on Graph Structure Based Method for Chinese Text Representation [J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(4): 618-624.)

[2] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[3] 何力，贾焰，韩伟红，等. 大规模层次分类问题研究及其进展[J]. 计算机学报，2012，35(10): 2101-2115. (He Li, Jia Yan, Han Weihong, et al. Research and Development of Large Scale Hierarchical Classification Problem [J]. Chinese Journal of Computers, 2012, 35(10): 2101-2115.)

[4] Silla C N, Freitas A A. A Survey of Hierarchical Classification Across Different Application Domains [J]. Data Mining and Knowledge Discovery, 2011, 22(1-2): 31-72.

[5] Zhang C, Xue G R, Yu Y, et al. Web-scale Classification with Naive Bayes [C]. In: Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain. 2009.

[6] Medelyan O, Milne D, Legg C, et al. Mining Meaning from Wikipedia [J]. International Journal of Human-Computer Studies, 2009, 67(9): 716-754.

[7] Muchnik L, Itzhack R, Solomon S, et al. Self-emergence of Knowledge Trees: Extraction of the Wikipedia Hierarchies [J]. Physical Review E, 2007, 76(1): 1-12. DOI: http://dx.doi.org/10.1103/PhysRevE.76.016106.

[8] Gabrilovich E, Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis [C]. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. 2007: 1606-1611.

[9] Aggarwal N, Asooja K, Buitelaar P. Exploring ESA to Improve Word Relatedness [C]. In: Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics. 2014: 51-56.

[10] Milne D N, Witten I H, Nichols D M. et al. A Knowledge-Based Search Engine Powered by Wikipedia [C]. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. 2007.

[11] Chakrabarti D, Mehta R. The Paths More Taken: Matching DOM Trees to Search Logs for Accurate Webpage Clustering [C]. In: Proceedings of the 19th International Conference on World Wide Web. 2010.

利益冲突声明:

作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 夏天. wiki6629.zip. 由 6629 条维基百科文章构成的测试数据集，采用 XML 格式，每篇文章包含标题、截取的 300 字左右的正文、隶属的类别信息.

[2] 夏天. generated_paths.zip. 在测试集基础上，增加了利用本文方法生成的语义路径信息.

[3] 夏天. data_code_url.txt. 维基百科原始数据集的下载链接地址和代码链接地址.

收稿日期: 2015-11-16
收修改稿日期: 2015-12-21

ChinaXiv合作期刊

Generating Hierarchical Paths of Chinese Text from Wikipedia

Xia Tian
(Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education,
Renmin University of China, Beijing 100872, China)
(School of Information Resource Management, Renmin University of China, Beijing 100872, China)

Abstract: [**Objective**] Generate hierarchical semantic paths of texts from Wikipedia. [**Methods**] We first establish article concept vector of Chinese texts from Wikipedia through explicit semantic analysis. And then, we mapped the vector to the category nodes of hierarchical-tree-like graph. Finally, we generated the hierarchical paths with the help of seed node information diffusion and top-down path selection, as well as optimization technology. [**Results**] The average relevance degree of the first generated hierarchical path was 54.10% on the test dataset, and the top 20 paths were sorted by relevance in the descending order. [**Limitations**] We did not analyze the effect of using different numbers of explicit concept vector to the quality of the generated path. [**Conclusions**] The hierarchical paths generated from Wikipedia can reflect the main semantic meaning of the given texts.

Keywords: Semantic path　Explicit semantic analysis　Hierarchical classification　Wikipedia

chinaXiv:201711.01237v1

Summon 发现服务开始提供 Altmetric 信息

ProQuest 子公司 Ex Libris 于近日宣布已集成 Altmetrics 到 Summon 发现服务之中, 极大地丰富了用户体验, 改进了内容发现。这是 ProQuest 和 Altmetric 之间共同合作的成果, 使得研究人员只需点击鼠标, 就能获悉一项研究成果的在线分享、评论和讨论情况。

图书馆开启 Summon 发现服务的 Altmetric 集成功能, Summon 发现服务中会显示一个 Altmetric 徽章。用户可以单击这个徽章来探索一条搜索结果(如文章)的相关讨论信息。这些信息由 Altmetric 公司从多个来源获取而来, 包括: 主流媒体、维基百科、博客、社交网络、参考咨询管理人员、出版后的同行评议论坛, 以及其他在线社区。

谈到这次的整合, Ex Libris 负责发现和交付解决方案的副总裁 Shlomi Kringel 认为: “通过增加学术内容的曝光率和提高搜索结果的价值来改进用户的研究体验, 对我们所有的服务来说都是一个重要的目标。将 Altmetric 徽章加入 Summon 发现服务, 使得我们的用户能够更容易判断一项研究成果在学术界和读者中的影响力, 以及产生这一影响力背后的原因。”

Altmetric 公司创始人 Euan Adie 补充道: “我们很高兴看到 ProQuest 将 Altmetric 集成到了 Summon 发现服务之中。我们希望, 在与研究成果相关的在线活动被更多用户看到的同时, 用户也能更积极地参与到各自领域正在进行的有关学术成果的讨论之中。”

无需订阅 Altmetric.com, 图书馆就可以激活 Altmetric 徽章, 这样, Altmetric 徽章将显示在所有通过 ProQuest 平台, 如 360 Links、Ex Libris Primo 以及 Summon 发现服务等提供的搜索结果中。

(编译自: http://www.proquest.com/about/news/2016/Altmetric-data-now-available-in-the-Summon-Discovery-Service.html)

(本刊讯)